Pacific Northwest
NATIONAL LABORATORY

PNNL-34811

Gumby

Quantifying multi-modal model resiliency

September 2023

Eleanor Byler
Elise Bishoff
Charlie Godfrey
Myles McKay

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-34811

Gumby

Quantifying multi-modal model resiliency

September 2023

Eleanor Byler
Elise Bishoff
Charlie Godfrey
Myles McKay

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

With the rise of cheap data and sensors, more use cases are emerging for multi-input models. Research has shown that including multiple data modalities can improve performance, suggesting that deep learning models can successfully learn to leverage complementary information from different modalities. However, this improved predictive power comes with unanticipated costs: additional inputs change model resiliency and expand the threat space for adversarial attacks. We first provide theoretical underpinnings for how adversarial success scales with input dimension. We then characterize the performance of a suite of multispectral deep learning models with different fusion approaches, quantify their relative reliance on different input bands, and evaluate their robustness to naturalistic and adversarial image corruptions.

Acknowledgments

This research was supported by the **National Security Mission Seed**, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). The computational work was performed using PNNL Computing at Pacific Northwest National Laboratory. PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Contents

Figures

Tables

1.0 Introduction

Over the past decade, there has been growing interest in understanding the robustness of deep learning models. Robustness refers to a model's ability to maintain performance under various input shifts, including natural shifts (e.g., weather, environment) and adversarial shifts (e.g., attacks or digital perturbations). While many advancements have emerged in the field of robustness, deep learning models remain vulnerable to various attacks and distribution shifts. To date, much of this research has focused on evaluating model performance on image classification tasks using benchmark RGB image datasets. As such, our understanding of model robustness for other tasks and data modalities remains incomplete.

Additional information may improve a model's ability to distinguish malicious inputs or simply provide new attack avenues and vulnerabilities. In Section 2.0 (work published in [1]), we investigate how adversarial vulnerability depends on the dimension of the space of model inputs. Since incorporating data from additional sensors in model inputs necessarily increases the input dimensionality, the results obtained from the study provide a means of estimating the risk of expanding the amount of information included in model input data.

In Section 3.0 (work published in [3], [7]) we consider both adversarial robustness and natural robustness for deep vision models applied to overhead multispectral imagery, focusing on the combination of RGB and near-infrared (NIR) bands. In [3], we quantify the robustness of different model architectures and their relative reliance on different inputs in classification and segmentation tasks. In [7], we consider both adversarial robustness and natural robustness for multispectral segmentation models, placing an emphasis on perturbations and attacks that are physically meaningful.

2.0 Quantifying the Relationship Between Adversarial Vulnerability and Input Dimension

Adversarial examples are data points that have been intentionally been modified in some (often imperceptible to humans) way with the goal of causing a machine learning model to output an incorrect prediction. Vulnerability to adversarial examples as often viewed as an obstacle to deploying deep learning systems, especially in situations where the stakes of incorrect prediction are relatively high. Since they were first identified in [2] a vast body of empirical work has shown that essentially all neural networks are susceptible to adversarial examples, and there has been a strong sense that a model's level of vulnerability is strongly connected to the dimension of its input space. This connection has been mined by a range of works which use it as a perspective with which to explain the prevalence of adversarial examples in certain model types (e.g., computer vision). As deep learning models are applied to more and more safety critical applications, there is also an increasing practical relevance to understanding any general connections between adversarial vulnerability and the properties of a problem. In such settings, a simple statistic that can be easily computed (such as model input dimension) is useful for gauging the general adversarial risk for a proposed deep learning system.

This is especially true when the proposed system uses less familiar modalities/tasks to which one cannot easily refer to studies in the literature. For example, suppose one needs to evaluate the safety of applying deep learning to the output of a range of different sensors. Past work has considered the ambient dimension in which this data is collected. Should we worry less if a sensor captures a signal as a 50-dimensional vector rather than a 5,000-dimensional vector? In this paper we take this line of reasoning a step further and ask how this situation changes when instead of changing the ambient dimension we change the dimension of the subspace in which an adversary is constrained to perturb input. Such a thought experiment has practical relevance. Suppose that of the 500 input dimensions to our model, we believe that an adversary is only likely to get access to 50 dimensions (this may happen in multimodal settings where an adversary has much better access to a subset of the modalities). How should we compare this to a situation in which are only able to perturb a fixed 100-dimensional subspace of the input? How about a 5-dimensional subspace?

Motivated by this, in this work we revisit the connection between dimension and adversarial vulnerability. Unlike most other works in this research area, which look at susceptibility to adversarial examples as a function of the number of input dimensions dim(X) alone, we explore model susceptibility to adversarial examples constrained to a subspace $V \subseteq X$ as a function of dim(V)/dim(X). We find that unsurprisingly, for fixed dim(X), as dim(V) decreases average adversarial success rate (ASR) also decreases, though ASR only drops significantly when the quotient dim(V)/dim(X) drops below around 10% (see figure below). In other words, a model remains vulnerable when an adversary is only able to perturb a subset of input dimensions, but as this subset covers an ever smaller fraction of the available dimensions an adversary has to put increasing effort into finding adversarial examples.

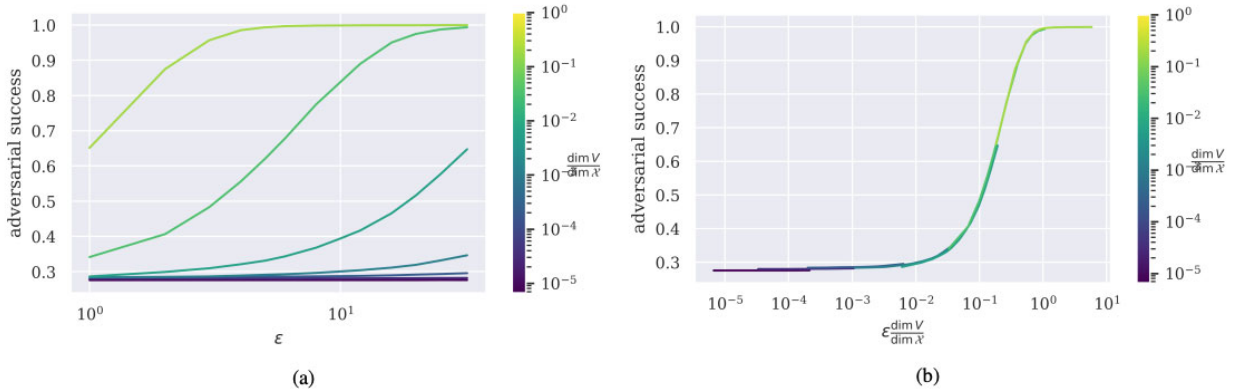

Figure 1: (a) Success of PGD adversarial attacks on an ImageNet trained ResNet50 constrained to subspaces V ⊆ X spanned by dim V randomly selected standard basis vectors. Adversarial examples are computed for a random subsample of 10,000 datapoints from the ImageNet validations set. The x-axis is the bound ϵ used during example generation and the different colored curves indicate the dimension dim V of the subspace to which the examples were constrained to, relative to the dimension dim X (= $3 \cdot 224^2$) of the ambient input space. When only a small number of dimensions can be perturbed, adversarial examples are challenging to generate even with large ϵ-bounds. (b) These curves become almost perfectly aligned when we reparameterize the x-axis by scaling by dim(V)/dim(X).

We further study how the adversarial budget ϵ (the extent to which a model input can be modified) interacts with dim(V) and dim(X). We find that the relationships of ϵ to ASR for different dim(V) are nearly identical up to scaling: more specifically, suppose that $C_1: \mathbb{R} \to [0,1]$ and $C_2: \mathbb{R} \to [0,1]$ map adversarial budget to ASR when adversarial examples are constrained to subspaces V_1 and V_2 respectively. We find that

$$C_1\left(\left(\frac{\dim V_1}{\dim X}\right) \quad \epsilon\right) \approx C_2\left(\left(\frac{\dim V_2}{\dim X}\right) \quad \epsilon\right).$$

This points to a strong relationship between dim(V), p, and ϵ that to our knowledge is novel. It further tells us that risk from adversarial examples can be mitigated by either restricting the dimensions that data can be manipulated (dimV) in or restricting the amount they can be manipulated before they are noticed (ϵ). This relationship is consistent across values of dim(V)/dim(X): if one wanted to understand the risk of an adversary perturbing data in a 50-dimensional subspace of a 500-dimensional-input space, one could for example estimate the success rate of an adversary with access to the entire input space and extrapolate using our equation. Finally, we provide a theoretical backing for our results as well as analyze their implications on common theories behind the prevalence of adversarial examples.

3.0 Deep Learning Models Applied to Overhead Multispectral Imagery for Image Classification and Segmentation

With the wealth of publicly available satellite imagery data and the development of large annotated satellite imagery datasets, deep learning models routinely achieve state-of-the-art performance in a number of important remote sensing applications, including land cover classification, agricultural monitoring, and disaster assessment. Typically, satellite sensors collect multispectral imagery, or imagery observed at wavelength bands beyond the traditional Red, Green, and Blue (RGB) bands found in natural imagery datasets. For many overhead imagery applications, spectral bands beyond the visible spectrum (e.g., near-infrared or short-wave infrared) are essential in distinguishing different surface materials or penetrating atmospheric haze. Deep learning models that leverage multispectral imagery are becoming increasingly common, and outperform RGB-only models in some applications [8], [9], [10].

For a given model, there are many possible ways to synthesize or fuse information from different inputs or data modalities. Models that combine data modalities at the input stage are sometimes called “early fusion” models (e.g., a 4-band image, or projecting 3D LiDAR data onto an RGB image). In contrast, models that process the different data modalities separately and combine them after feature extraction or in the final layer before classification are called “late fusion” models. Here, we are specifically interested in quantifying the robustness of different fusion approaches. To this end, we explore different combinations of input bands (NIR, RGB, RGB+NIR) and architectures (early vs. late fusion) to better understand how each of these variables affects the model’s overall robustness, and explore any potential trade-offs with model performance in image classification and segmentation tasks.

For adversarial robustness, we focus on data poisoning, wherein adversaries inject malicious data into the training data to cause erroneous classifications or backdoor access during test time. For natural robustness, we develop an approach for physically realistic perturbations that can be applied to multispectral imagery, building upon ImageNet-C [11], the industry standard for RGB imagery. We then quantify the robustness of the segmentation models, assessing the relative accuracy and robustness of different fusion approaches as compared to an RGB model baseline.

In 3.1, we quantify the performance and robustness of models with different fusion approaches, and assess model reliance on different input types. In 3.2, we provide a more comprehensive picture of model robustness and consider both adversarial and natural robustness.

3.1 Architecture impact on robustness and interpretability

In [3], we study multispectral models operating on RGB and NIR channels on two different data sets and tasks (one involving image classification, the other image segmentation). In addition, for each dataset/task we consider two different multispectral fusion architectures, early and late (see Figure 2).

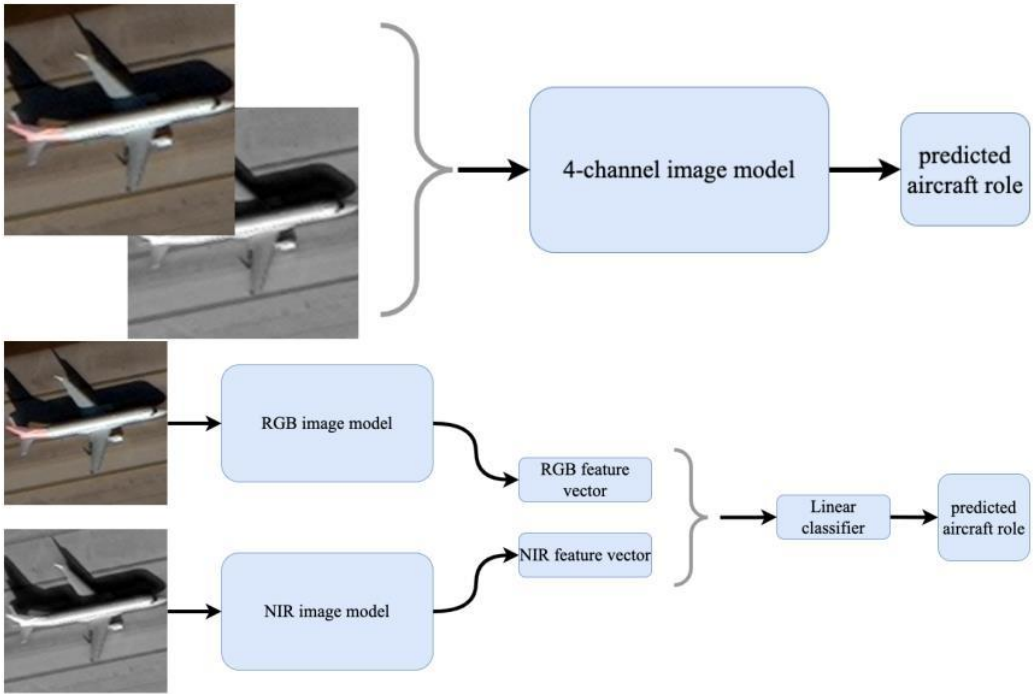

Figure 2: RGB+NIR fusion architectures for classifiers predicting aircraft role. Top: early, bottom: late. Braces denote image/feature concatenation.

Even when different fusion architectures achieve near-identical performance as measured by test accuracy, they leverage information from the various spectral bands to varying degrees: we find that for classification models trained on a dataset of RGB+NIR overhead images, late fusion models place far more importance on the NIR band in their predictions than their early fusion counterparts. Here we measure "importance" using a metric called *perceptual score* [4]. In contrast, for segmentation models we observe that both fusion styles resulted in models that place greater importance on RGB channels, and this effect is more pronounced for late fusion models. See Figure 3.

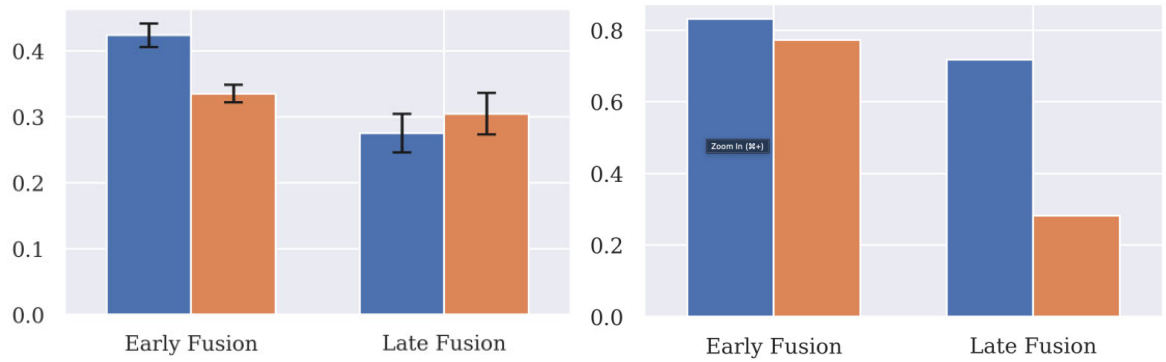

Figure 3: **Left:** perceptual scores for the multispectral classifiers on the RarePlanes dataset [5]. Blue: RGB, orange: NIR. The early fusion models have a higher perceptual score for RGB channels (i.e., more reliance on RGB inputs), whereas the late fusion models have higher perceptual score for NIR channels (i.e., more reliance on NIR input). **Right:** perceptual scores for the multispectral segmentation models on the US3D dataset [6]. Both early and late fusion models have higher perceptual scores for RGB data, demonstrating that model performance

relies more strongly on the RGB inputs. For late fusion models this effect is even more dramatic, suggesting that the NIR input is less important, in contrast to the classification model scores shown at left.

Perhaps unsurprisingly, these effects are mirrored in an evaluation of model robustness to naturalistic image corruptions affecting one or more input channels — in particular, early fusion classification models are more sensitive to corruptions of RGB inputs, and segmentation models with either architecture are comparatively immune to corruptions affecting NIR inputs alone (see Figure 4, Figure 5). In order to carry out this analysis, we created to the best of our knowledge the first benchmark datasets for evaluating robustness of multispectral image models to naturalistic corruptions (see Figure 9, Figure 10 for example datapoints).

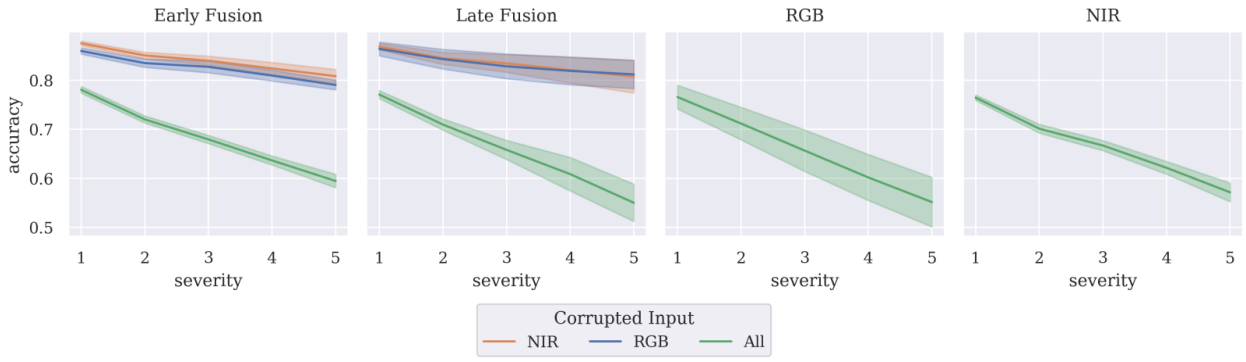

Figure 4: Corruption robustness of RarePlanes classifiers. Each subplot corresponds to a model architecture, and each line corresponds to a choice of input (RGB, NIR or both) to corrupt. Accuracy is averaged over 15 types of corruptions.

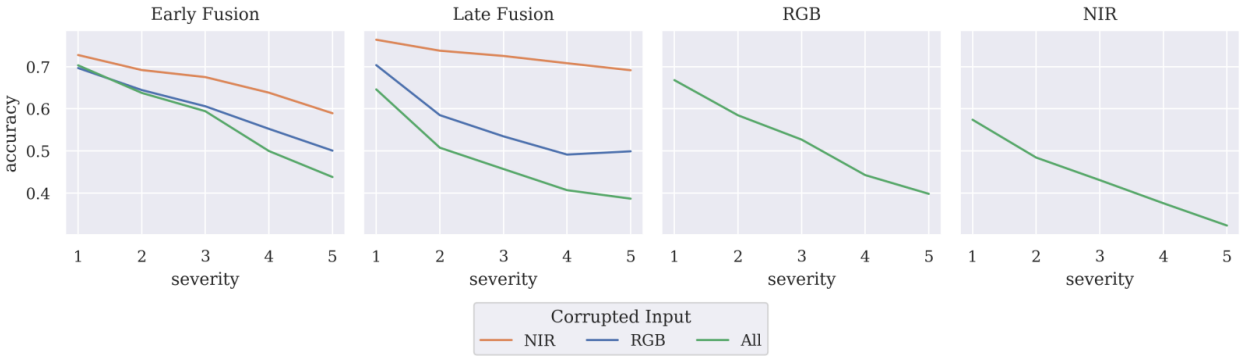

Figure 5: Corruption robustness of US3D segmentation models. Each subplot corresponds to a model architecture, and each line corresponds to a choice of input (RGB, NIR or both) to corrupt. IoU (a segmentation model accuracy metric) is averaged over 15 types of corruptions.

On the whole, our experiments suggest that segmentation models and classification models use multispectral information in different ways.

3.2 Robustness to natural and adversarial corruptions

In [7], we study the adversarial and natural robustness of multispectral classification and segmentation models for overhead imagery. While existing adversarial and natural robustness research has focused primarily on digital perturbations, we prioritize on creating realistic

perturbations designed with real-world physical conditions in mind. For adversarial robustness, we focus on data poisoning attacks, as shown in Figure 6. For natural robustness, we focus on extending ImageNet-C common corruptions [11] for fog and snow that coherently and self-consistently perturbs the input data, as shown in Figure 7.

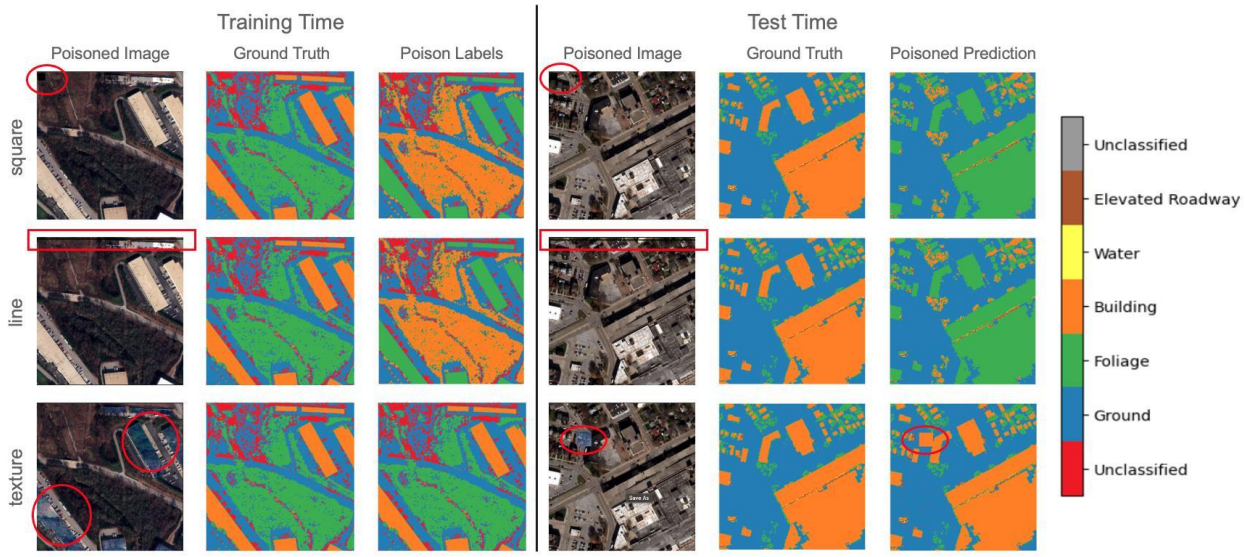

Figure 6: Examples of data poisoning attacks implemented in this work: square, line, and texture. The square and line attacks (top and middle rows) operate like a trigger; when present, the model should erroneously classify foliage pixels as the "building" class. In contrast, the texture attack (bottom row) trains the model to learn a targetable representation - here, foliage that is classified as a building. All attacks were highly successful with only 10% of the training data poisoned.

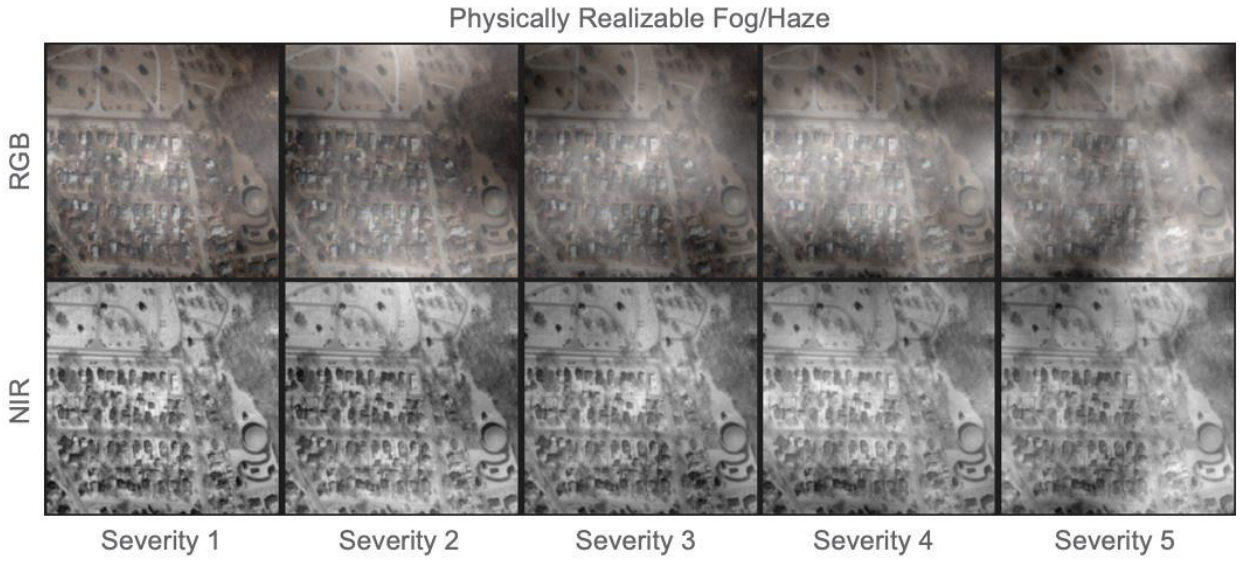

Figure 7: An example of the physically realistic fog/haze perturbations used in this work. We modify the original implementation of the ImageNet-C perturbations to account for the fact that NIR light more easily penetrates fog, haze, and smoke.

We compare the performance and robustness of the multispectral segmentation models for the two fusion approaches introduced in Section 3.1 (early, late) on the US3D dataset. Baseline performance metrics are shown in columns two and three of Table 1. All models perform similarly across the considered metrics (pixel accuracy and IOU score), with the multispectral early fusion model showing the best overall performance. The last column of Table 1 shows the adversarial robustness of the models, as measured by the success of the poisoning attacks. Overall, we find both RGB and multispectral models are vulnerable to data poisoning attacks regardless of input or fusion architectures, with adversarial success rates of over 90%. Interestingly, while the multispectral models are the best performing models, they also have the highest overall adversarial success rates. This suggests that the extra information provided by additional bands boosts performance but also reduces the overall adversarial robustness.

Table 1: Baseline Segmentation Model Performance and Adversarial Robustness

Model Input	Pixel Accuracy	IOU Score	Adversarial Robustness Poisoning Success Rate
NIR	0.919	0.757	**0.921**
RGB	0.917	0.761	0.924
Early Fusion	**0.921**	**0.769**	0.932
Late Fusion	0.917	0.764	0.937

For natural robustness, we show the performance degradation from natural perturbations in Figure 8. We find that the physically realizable natural perturbations degrade model performance in all cases, however the impact differs with fusion architecture and input data. In particular, the early and late fusion models show improved robustness to natural perturbations over the baseline RGB model, suggesting that these models are able to leverage information from additional bands to improve segmentation performance in adverse weather conditions. The early fusion model shows the best overall robustness; which agrees with findings in Section 3.1 that suggest that early fusion models rely more on NIR inputs.

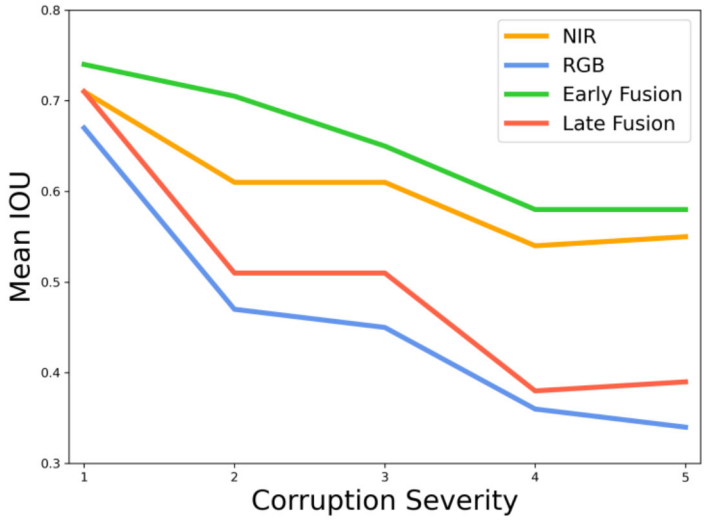

Figure 8: **Natural Robustness**. Segmentation model performance on data corrupted with physically realistic snow at varying levels of severity (see example datapoints in Figure 7).

Our findings can be summarized accordingly:

1. We find that all segmentation models are vulnerable to data poisoning attacks, regardless of input (NIR, RGB, NIR+RGB) or fusion architecture (early, late). Both the fine-grained attacks and physically realizable texture attacks are highly successful (ASR > 90%) with only 10% of the training images poisoned; however, the texture attacks are less likely to be detected by the victim.
2. The two RGB+NIR models show the best overall performance as measured by accuracy and IOU, but also the worst overall robustness to adversarial attacks. We conclude that the additional information provided by the additional input bands boosts overall performance, but does so at the expense of adversarial robustness.
3. In contrast with previous work in object detection [12], we did not find any significant difference in adversarial robustness between early and late fusion approaches, suggesting that the adversarial robustness of fusion approaches varies with attack type and model task.
4. We create a physically realistic version of the ImageNet-C snow and fog corruptions that are appropriate for multispectral data and faithfully preserve the real-world observational signatures of snow and fog/haze.
5. We find that both RGB+NIR models show improved robustness to natural perturbations over RGB only models, suggesting that these models are able to successfully leverage NIR information to improve segmentation performance in adverse weather conditions. We find that the early fusion models have the best overall natural robustness, which aligns with our results from [3] that find that the early fusion models rely more on NIR inputs. Additionally, the foliage class, which has a distinct NIR signature, shows significant improvement in the early fusion model.

4.0 References

[1] Charles Godfrey, Henry Kvinge, Elise Bishoff, Myles McKay, Davis Brown, Tim Doster, Eleanor Byler, “How many dimensions are required to find an adversarial example?”; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 2352-2359; https://arxiv.org/abs/2303.14173

[2] C. Szegedy *et al.*, “Intriguing properties of neural networks,” *arXiv:1312.6199 [cs]*, Feb. 2014. Available: http://arxiv.org/abs/1312.6199

[3] Charles Godfrey, Elise Bishoff, Myles McKay, Eleanor Byler, "Impact of model architecture on robustness and interpretability of multispectral deep learning models," Proc. SPIE 12519, Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imaging XXIX , 125190L (13 June 2023); https://doi.org/10.1117/12.2662998

[4] I. Gat, I. Schwartz, and A. Schwing, “Perceptual Score: What Data Modalities Does Your Model Perceive?,” *arXiv:2110.14375 [cs]*, Oct. 2021. Available: http://arxiv.org/abs/2110.14375

[5] J. Shermeyer, T. Hossler, A. Van Etten, D. Hogan, R. Lewis, and D. Kim, “RarePlanes: Synthetic Data Takes Flight,” *arXiv:2006.02963 [cs]*, Nov. 2020. Available: http://arxiv.org/abs/2006.02963

[6] B. Le Saux, N. Yokoya, R. Haensch, and M. Brown, “2019 IEEE GRSS Data Fusion Contest: Large-Scale Semantic 3D Reconstruction [Technical Committees],” *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 4, pp. 33–36, Dec. 2019, doi: 10.1109/MGRS.2019.2949679.

[7] Elise Bishoff, Charles Godfrey, Myles McKay, Eleanor Byler, "Quantifying the robustness of deep multispectral segmentation models against natural perturbations and data poisoning," Proc. SPIE 12519, Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imaging XXIX , 125190M (13 June 2023); https://doi.org/10.1117/12.2663498

[8] Tian, M., Ban, S., Yuan, T., Ji, Y., Ma, C., and Li, L., “Assessing rice lodging using UAV visible and multispectral image,” 42(23), 8840–8857. Publisher: Taylor & Francis. eprint: https://doi.org/10.1080/01431161.2021.1942575.

[9] Xu, X., Li, Y., Wu, G., and Luo, J., “Multi-modal deep feature learning for RGB-d object detection,” Pattern Recognition, Volume 72, 2017, pp. 300–313. https://doi.org/10.1016/j.patcog.2017.07.026

[10] Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., and Burgard, W., “Multimodal deep learning for robust RGB-d object recognition.” Number: arXiv:1507.06821.

[11] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D., “Natural adversarial examples.”

[12] Wang, S., Wu, T., Chakrabarti, A., and Vorobeychik, Y., “Adversarial robustness of deep sensor fusion models.” Number: arXiv:2006.13192.

Appendix A – Datasets for Benchmarking Robustness of Multispectral Image Models

In Figure 9 and Figure 10, we provide example datapoints obtained by applying naturalistic corruptions to images from RarePlanes.

Figure 9: RGB corruptions of a RarePlane chip from our test set.

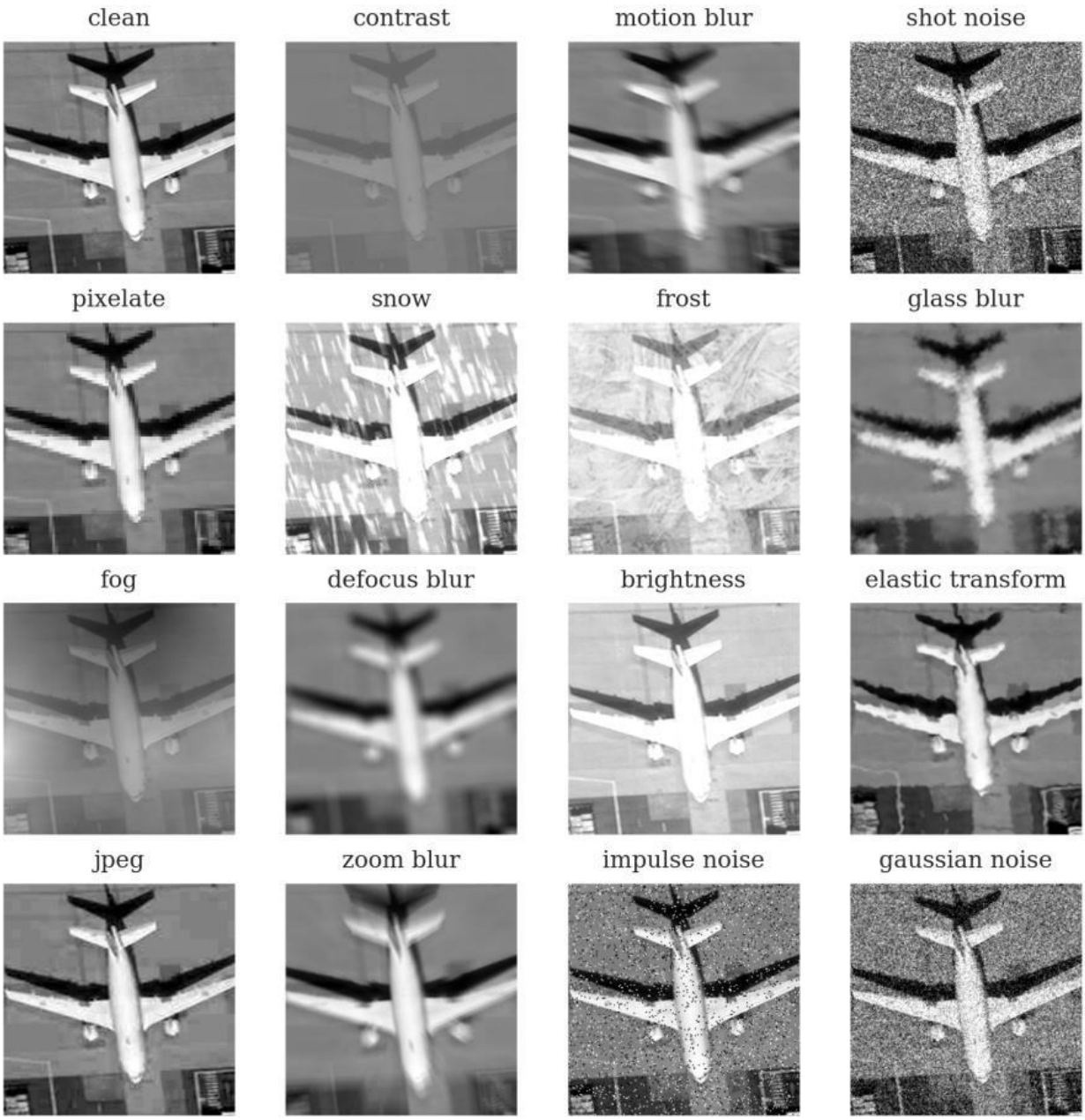

Figure 10: NIR corruptions of a RarePlane chip from our test. Note that the motion blur (2nd row, 3rd column) is applied in the same direction as in Figure 9.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov